논문 2008-45IE-3-4

잡음 섞인 한국어 인식을 위한 ICA 비교 연구

(Comparison of ICA Methods for the Recognition of Corrupted Korean Speech)

김 선 일*

(Seonil Kim)

요 약

두 가지 Independent Component Analysis(ICA) 알고리즘을 적용하여 자동차 엔진 소음과 섞인 음성 신호의 인식을 시도하였다. 이를 이용하여 추정한 신호를 HMM을 이용하여 인식하였고 이 신호의 인식률을 소음이 섞이기 전의 음성 신호의 인식률과 비교하였다. 음성 신호를 추정하는데 두 가지 서로 다른 ICA를 사용하였으며 그 중의 하나는 negentropy 를 최대화하는 FastICA 알고리즘이며 다른 하나는 출력 신호 사이의 독립성을 최대화하여서 입력과 출력 사이의 mutual information을 최대화하는 information-maximization approach 이다. 남성 앵커가 진행한 한국어 뉴스 문장에 대한 단어 인식률은 87.85%이며 다양한 신호 대 잡음비를 갖도록 소음을 섞어서 추정을 한 후 인식을 시도한 결과 FastICA를 이용해 추정한 음성 신호에 대한 인식률은 1.65%, information-maximization을 이용해 추정한 음성 신호에 대한 인식률은 2.02% 인식률 저하가 나타났다. 따라서 어느 방법을 적용하든지 의미 있는 차이가 없음을 확인하였다.

Abstract

Two independent component analysis(ICA) algorithms were applied for the recognition of speech signals corrupted by a car engine noise. Speech recognition was performed by hidden markov model(HMM) for the estimated signals and recognition rates were compared with those of original speech signals which are not corrupted. Two different ICA methods were applied for the estimation of speech signals, one of which is FastICA algorithm that maximizes negentropy, the other is information-maximization approach that maximizes the mutual information between inputs and outputs to give maximum independence among outputs. Word recognition rate for the Korean news sentences spoken by a male anchor is 87.85%, while there is 1.65% drop of performance on the average for the estimated speech signals by FastICA and 2.02% by information-maximization for the various signal to noise ratio(SNR). There is little difference between the methods.

Keywords : ICA, HMM, Negentropy Mutual Information, Maximum Independence, Car Engine Sound, Speech Recognition

I. Introduction

Although the performance of speech recognition technology has advanced over a few decades, speech recognition in a real world has many problems to tackle. One of those is all kinds of sounds from other sources being called noise. Especially if you want to order your car to open windows while you are driving, such noise as the sound from the engine of your car is very critical.

The interaction between human and computer is increasingly important in today's technological society. In the area of telematics which needs to respond to the voices from driver, speech recognition technology is very important. But the interfering sounds from car engine noise, music and wind make the performance of speech recognition unacceptable

Telematics incorporates several technologies to serve people while they drive. Among those technologies speech recognition is. To recognize

* 정회원, 거제대학 조선정보기술계열
(Department of Information Technology for Shipbuilding, Koje College)
접수일자: 2008년6월30일, 수정완료일: 2008년8월7일

speech signals corrupted with noises you have to first recover them from the mixed ones. ICA can be one of the several methods that can recover speech signals.

ICA uses such tools to recover speech signals as kurtosis[1] and negentropy[2] to get the independent source signals from the observed signals by maximizing nongaussianity. For the information theory, information maximization[3] between observed signals and estimated signals, and minimization of mutual information[4~5] between estimated signals based on the self organizing learning algorithm are being used. This is conventional approach to get the solution.

Hyvarinen[6] proposed fast fixed-point algorithm known as FastICA. It is very fast. But is it superior to the other methods for the Korean speech signals.

Such question can be answered by the comparison of two ICA methods for the recognition of corrupted Korean speech which reveals at the end of this paper. If there is little difference between two methods that can be approached in totally different ways, you can freely choose the method you want according to your environment.

First, blind source separation(BSS) is explored briefly. Secondly, the ICAs which include FastICA algorithm[6] and an information-maximization approach[3] are explained. Lastly, the results of speech recognition experiments for the original and estimated source signals are shown.

Ⅱ. BSS

BSS is the techniques to get the solution for the segregation of unobserved signals or sources from several observed mixtures[7]. Using BSS enhancement of speech signals in a noisy car environment can be achieved[8]. BSS can be formulated mathematically as the estimation of m latent signals from their n mixed signals. That is

$$\mathbf{x}=\mathbf{A}\mathbf{s} \tag{1}$$

where

$$\mathbf{x}=[x_1\, x_2\, x_3 \cdots x_n]^{\mathrm{T}} \tag{2}$$

$$\mathbf{s}=[s_1\, s_2\, s_3 \cdots s_m]^{\mathrm{T}} \tag{3}$$

$\mathbf{x}$ is a column vector which consists of $\mathbf{n}$ observed signals. x_i is the output signal of i^{th} microphone or sensor which mixes m signals which are as independent as possible[9].

$\mathbf{s}$ is a column vector which consists of m mutually independent source signals and $\mathbf{A}$ is a mixing matrix.

If the number of microphones or sensors is the same as the number of independent signals, $n=m$.

To recover source signals in $\mathbf{s}$ from the mixed signals in $\mathbf{x}$, you have to know mixing matrix $\mathbf{A}$.

Then $\mathbf{s}$ can be found as

$$\mathbf{s}=\mathbf{A}^{-1}\mathbf{x}=\mathbf{W}\mathbf{x} \tag{4}$$

$$\mathbf{W}=\mathbf{A}^{-1} \tag{5}$$

Since mixing matrix $\mathbf{A}$ is not known, $\mathbf{W}=\mathbf{A}^{-1}$ can't be found. It has to be estimated from the mixed signals which can be observed to find source signals. So the term blind is used. To recover original signals $\mathbf{W}$ which gives maximum independence between outputs in $\mathbf{x}$ has to be estimated.

Ⅲ. ICA

ICA assumes the maximum independence between source signals which enables estimation.

There are several approaches to achieve ICA and get estimation of $\mathbf{W}$. Maximization of nongaussianity, maximum likelihood estimation and minimization of mutual information are those[2]. They are based on statistical inference and information theory[10].

It is assumed that source signals are nongaussian. Fortunately, speech signals are known usually to have supergaussian probability distribution which is nongaussian[1].

The central limit theorem is a classic result in probability theory. It says that the distribution of sum of independent random variables tends toward a gaussian distribution. A sum of two independent

random variables usually has a distribution that is closer to gaussian than any of the two original random variables. It means that nongaussianity is independent. W which gives maximum nongaussinity between outputs can be estimated..

How can it be measured? It can be measured by kurtosis[1]. The kurtosis of a random variable y can be defined as

$$kurt(y) = E\{y^4\} - 3(E\{y^2\})^2 \quad (6)$$

To simplify things, They are made to be sphere so that

$$E\{y^2\} = 1 \quad (7)$$

$$E\{y\} = 0 \quad (8)$$

Then (8) becomes

$$kurt(y) = E\{y^4\} - 3 \quad (9)$$

For a gaussian distribution, kurtosis becomes 0. For most nongaussian random variables, kurtosis is nonzero.

It's very simple. You do calculate only $E\{y^4\}$. But it has drawback that higher order statistics is sensitive to peak noise. Accordingly, kurtosis is not a robust measure of nongaussianity

1. Whitening

Independence means uncorrelatedness. If the covariance of random variables is zero, they are uncorrelated. But uncorrelatedness does not imply independence[11].

Whiteness means that their mean is zero and their covariance is 1. It is slightly stronger property than uncorrelatedness. Whitening means independence between vectors. But it is not sufficient. It is the preprocessing to solve the problem of ICA. They are made to be white before proceeding to the algorithm of ICA.

It is assumed that z is whitened vector of x which consists of mixed signals of speech and engine noise of a car(called engine noise).

$$\mathbf{z} = \mathbf{V}\mathbf{x} \quad (10)$$

where V is a whitening matrix. Observed data vector x is transformed to whitened one by linearly multiplying V. The whitening matrix can be found as

$$\mathbf{V} = \mathbf{E}\mathbf{D}^{-\frac{1}{2}}\mathbf{E}^{\mathrm{T}} \quad (11)$$

where

E is the orthogonal matrix of eigenvectors of covariance matrix of x,

D is diagonal matrix of the eigenvalues of covariance matrix of x.

Accordingly, all means of z are 0s and covariance matrix of z is the identity matrix.

2. FastICA algorithm using negentropy

Negentropy is another measure of nongaussianity. It is based on the information theory. It can be avoided to be sensitive to peak noise.

The entropy of random vector y is

$$H(y) = -\int p_y(\eta)\log p_y(\eta)d\eta \quad (12)$$

The definition of negentropy J of random vector y is given by

$$J(\mathrm{y}) = \mathrm{H}(\mathrm{y}_{gauss}) - \mathrm{H}(\mathrm{y}) \quad (13)$$

where y_{gauss} is a gaussian random vector. It has the same covariance as random vector y.

Gaussian random vector y has a hightest entropy, negentropy becomes 0 when y has a gaussian distribution. Otherwise, negentropy becomes bigger than 0. Bigger negentropy, less gaussian. Negentropy can be a measure of nongaussianity.

It is computationally too expensive to get the negentropy using equation (13). They can be reduced by good approximations.

$$J(y) \approx \frac{1}{12}E\{y^3\}^2 + \frac{1}{48}kurt(y)^2 \quad (14)$$

More approximations give

$$J(y) \propto [E\{G(y)\} - E\{G(\nu)\}]^2 \quad (15)$$

where ν is a gaussian variable of zero mean and unit variable. The random variable y is assumed to have zero mean and unit variance. By appropriate function G you can get better negentropy than the one given by (15).

To measure by these equations you can use gradient method which takes the gradient of the approximation of negentropy in (15) with respect to the inversion of mixing matrix, $\mathbf{W}=\mathbf{A}^{-1}$. When the gradient approaches 0, you can get the good estimation of source signals.

Since $\mathbf{W}=(\mathbf{w}_1\,\mathbf{w}_2\cdots\mathbf{w}_m)^{\mathrm{T}}$ an independent source signal using whitened vector $\mathbf{z}$ can be estimated as

$$y=\mathbf{w}^{\mathrm{T}}\mathbf{z} \tag{16}$$

Fast fixed-point algorithm was used which was known as FastICA[6]. It is much faster than gradient method.

Fast fixed-point algorithm is based on Newton method which is described as

$$x_{n+1}=x_n-f(x_n)/f'(x_n) \tag{17}$$

where $f(x_n)$ is a function of x_n and $f'(x_n)$ is a derivative of $f(x_n)$.

According to the Lagrange conditions under the constraint $E\{y^2\}=1$, $\|\mathbf{w}\|=1$, the following equation is obtained.

$$f(\mathbf{w})=\mathrm{E}\{\mathbf{z}g(\mathbf{w}^{\mathrm{T}}\mathbf{z})\}+\beta\mathbf{w} \tag{18}$$

Its derivative comes to

$$f'(\mathbf{w})=\mathrm{E}\{\mathbf{z}\mathbf{z}^{\mathrm{T}}g'(\mathbf{w}^{\mathrm{T}}\mathbf{z})\}+\beta\mathrm{I} \tag{19}$$

Substituting (18) and (19) into (17),

$$\mathbf{w}\leftarrow E\{\mathbf{z}g(\mathbf{w}^{\mathrm{T}}\mathbf{z})-E\{g'(\mathbf{w}^{\mathrm{T}}\mathbf{z})\}\mathbf{w}\} \tag{20}$$

One of the following equations can be used as an good approximation of g.

$$g_1(y)=\tanh(a_1 y) \tag{21}$$

$$g_2(y)=y\exp(-y^2/2) \tag{22}$$

$$g_3(y)=y^3 \tag{23}$$

where $1\le a_1\le 2$.

You can update $\mathbf{w}$ according to (20). After normalizing $\mathbf{w}$, you can repeat (20) if it is not converged.

3. Information Maximization

A self-organizing learning algorithm performs maximization of the information transferred in a network of nonlinear units[3].

The information that output Y about input X contains is defined as

$$I(Y,X)=H(Y)-H(Y|X) \tag{24}$$

where $H(Y)$ is the entropy of the output. $H(Y|X)$ is whatever entropy the output has which didn't come from the input. That is $H(Y|X)=H(N)$. In the case that there is no noise, the mapping between X and Y is deterministic and the mutual information can be maximized by maximizing the entropy $H(Y)$ alone.

When a single input x is passed through a transforming function $y=g(x)$ to give an output variable y, both $I(y,x)$ and $H(y)$ are maximized when high density parts of the probability density function of x is aligned with highly sloping parts of the function $y=g(x)$.

When $g(x)$ is monotonically increasing or decreasing, the pdf of the output $f_y(y)$ can be written as a function of the pdf of the input $f_x(x)$.

$$f_y(y)=\frac{f_x(x)}{|\partial y/\partial x|} \tag{25}$$

The entropy of output is given by

$$H(y)=-E[\ln f_y(y)] \tag{26}$$

Substituting (25) into (26) gives

$$H(y)=E\left[\ln\left|\frac{\partial y}{\partial x}\right|\right]-E[\ln f_x(x)] \tag{27}$$

The second term on the right is just input, and not related to a parameter w, it can be neglected.

If $y=g(x)$ is the logistic transfer function

$$y=\frac{1}{1+e^{-wx}} \tag{28}$$

then the following can be derived

$$\Delta w \propto \frac{\partial H}{\partial w}=\frac{1}{w}+x(1-2y) \tag{29}$$

Extending to the input vector x

$$\Delta \mathbf{W} \propto [\mathbf{W}^{\mathrm{T}}]^{-1}+(1-2\mathbf{y})\mathbf{x}^{\mathrm{T}} \tag{30}$$

With the some modification not to calculate costly inverse calculation and learning rate, (30) can be used to learn weight matrix W.

IV. Recognition using Estimated Speech Signals

Speech recognition experiments were conducted for the Korean news corpus which was told by a male anchor. They were sampled at 16 bits, at a rate of 16 ㎑. The size of a frame is 25 ㎳, that is, 400 samples are there. This frame is moving by 10㎳ apart. Hamming window of preamplification coefficient of 0.97 was used. 26 coefficients are used as a feature vector, which includes 12 Mel frequency cepstrum coefficients(MFCC), 1 energy, and their differences. 3 state left-right monophone hidden Markov model(HMM) was used for the training and recognition. Used were 581 sentences totally. For the HMM training, 500 sentences were used. 81 sentences were put to work for the recognition. In other words, 5972 words were given to training. 880 words were applied for recognition. 10 gaussian mixtures were preferred for these experiments, since they showed best performance. Monophone model was adopted as data was insufficient for the triphone model training.

87.85% is the recognition rate for the words of sentences of original speech data(*Corr*), and the *Acc* is 85.77%.

You can see the recognition rates for the signals estimated by FastICA in Table 1, and by information-maximization in Table 2. The abbreviations are

표 1. Negentropy를 이용한 FastICA로 추정한 음성신호의 인식률

Table 1. Recognition rates for the estimated signals by the FastICA algorithm using negentropy.

SNR	Corr(%)		Acc(%)	
	1st	2nd	1st	2nd
0dB	86.70	86.70	85.11	85.23
-1dB	86.70	86.70	85.11	85.23
-2dB	86.59	86.59	85.11	85.00
-3dB	86.59	86.70	85.11	85.23
-4dB	86.82	86.48	85.23	85.00
-5dB	86.59	86.59	85.00	85.11
-6dB	86.48	86.52	85.00	85.34
-7dB	86.70	86.70	85.11	85.11
-8dB	86.48	86.93	85.00	85.34
-20dB	86.59	86.59	84.99	85.11
-25dB	85.45	86.25	83.64	84.20
-30dB	85.45	85.91	83.98	83.98
-35dB	84.77	85.45	83.18	83.86
-40dB	82.39	84.20	80.68	82.84
Mean	86.02	86.31	84.45	84.76
	86.20		84.61	

$$Corr = H/N * 100\% \tag{31}$$

$$Acc = (H-I)/N * 100\% \tag{32}$$

where N : The total number of labels,
H : The number of correct labels,
I : The number of insertions.

Two signals, a speech and an engine noise were mixed according to the signal to noise ratio(SNR) in table. Engine noise is measured sound from the muffler. It doesn't include other noises such as wind. The sound includes acceleration. Two mixed signals were made to simulate two microphones. Since it is assumed to have a distance between microphones. the difference between the energies of the two mixed signal was set to 2dB.

In tables, you can see recognition rates for the estimated speech signals being separated from mixed ones. Two experiments was performed twice for each method since the qualities of estimated speech signals are not even. And means were calculated. There is

표 2. Information-maximization을 이용해 추정한 음성 신호의 인식률

Table 2. Recognition rates for the estimated signals by an information-maximization approach.

SNR	Corr(%)		Acc(%)	
	1st	2nd	1st	2nd
0dB	86.02	85.91	84.43	84.09
-1dB	85.91	86.02	84.20	84.43
-2dB	85.57	85.00	84.09	83.41
-3dB	85.34	85.68	83.75	83.86
-4dB	86.02	85.45	84.32	83.64
-5dB	86.02	86.02	84.55	84.55
-6dB	86.25	85.80	84.66	84.20
-7dB	85.68	85.68	84.09	83.98
-8dB	86.25	85.57	85.00	83.98
-20dB	86.25	85.91	84.55	84.43
-25dB	85.57	86.36	84.09	84.89
-30dB	85.80	85.68	83.86	83.86
-35dB	85.57	86.36	83.86	84.89
-40dB	85.45	85.91	83.41	84.20
Mean	85.84	85.81	84.20	84.17
	85.83		84.19	

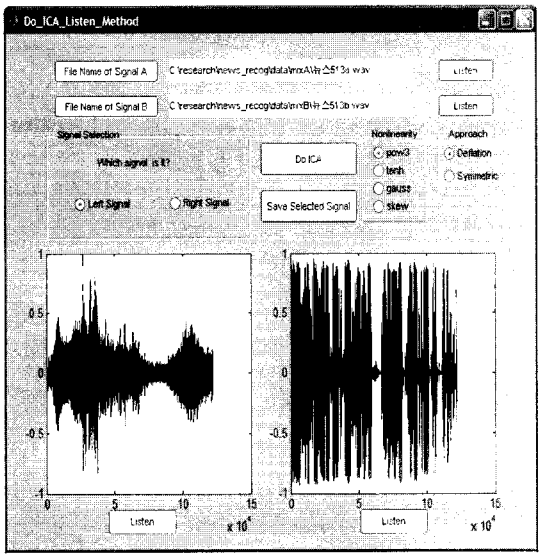
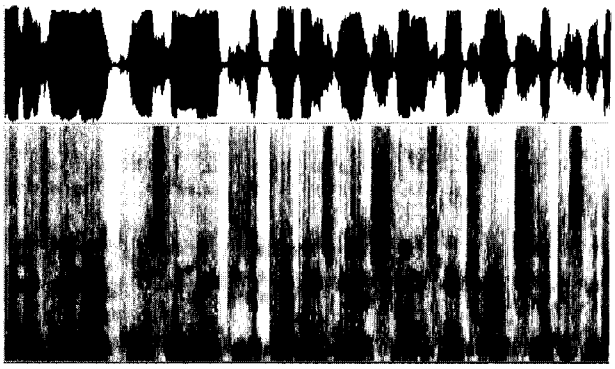


그림 1. 추정된 자동차 엔진 소음(왼쪽)과 음성 신호(오른쪽)

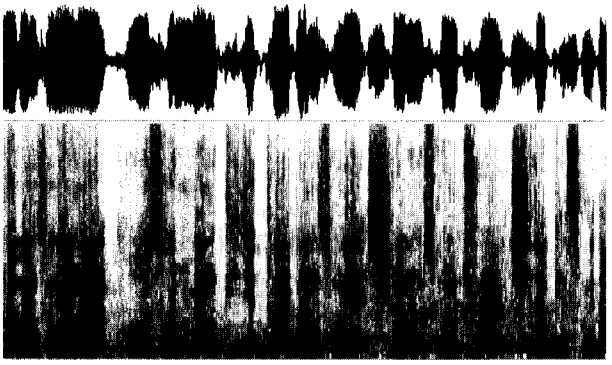
Fig. 1. Estimated car engine noise signal(Left) and speech signal(Right).

only 1.65% of degradation of *Corr* for the estimated speech signals using FastICA algorithm, 2.02% for those by information-maximization approach. As for *Acc*, 1.16% for those using FastICA algorithm, 1.58% for those using information-maximization approach.

The recognition rates for the estimated speech signals, even if only engine noise can be heard in



a) Spectrogram of source signal(speech).



b) Spectrogram of estimated speech signal.

그림 2. 음성 신호의 스펙트로그램

Fig. 2. Spectrogram of Speech Signal.

mixed signals(under -30dB), show good results. When the speech signal was degraded by an engine noise severely, the estimated speech signals include a little bit of engine noise you can notice. But the results show no disappointment. In Figure 1, estimated engine noise signal and speech signal are shown. They were estimated by the FastICA using deflation method and approximation function of (23).

Spectrograms for the speech signals are shown in Figure 2. Some noises can be found in the estimated speech signal in it, which came from the engine. But vocal tract is preserved as you can see in spectrogram. MFCC is sensitive in low frequency but insensitive in high frequencies like the ear of human. The noises present in the high frequency due to the lack of separation can be compensated by the feature vector MFCC.

FastICA using negentropy shows better results than the information-maximization in recognition rate. However The difference is not significant(0.37% *Corr*).

V. Conclusions

Speech signals were estimated from the ones mixed with engine noise signal. FastICA algorithm using negentropy and information-maximization approach were used to estimate speech signals from the mixed ones.

Two methods show reasonable results, which are acceptable for the application in real environments, although the experiments were performed in an artificial environment. The possibility of real application can be seen as the estimated speech signals by ICA are well recognized.

There may be unexpected problems to tackle in real environments. But it is true that at least the degradation of speech signals mixed with the engine noise is no longer obstacle for the speech recognition while driving.

There may be other ones such as the sound of wind which is passing by or the drops of rain on your car. These give another challenges to tackle.

Reference

[1] J. P. LdBlanc and P. L. De Leon, "Speech Separation by Kurtosis Maximization," Proc. ICASSP, vol. 2, pp. 1029-1032, 1998.

[2] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley and Sons, 2000.

[3] A. J. Bell land Terrence J. Sejnowski, "An information-maximisation approach to blind separation and blind deconvolution," Neural Computation, vol. 7, no. 6, pp. 1129-1159, 1995.

[4] P. Comon, "Independent component analys, A new concept?," Signal Processing, vol. 36, pp. 287-314, 1994.

[5] S. Amari and A. Cichocki, "A New Learning Algorithm for Blind Signal Separation," Advances in Neural Information Processing Systèm, vol. 8, pp. 757-763, MIT Press, 1996.

[6] A. Hyvarinen, "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis," IEEE Trans. On Neural Networks, vol. 10, no. 3, May, 1999.

[7] J. F. Cardoso, "Blind signal separation: statistical principles," Prod. IEEE, vol. 9, no. 10, pp. 2009-2025, Oct., 1998.

[8] E. Bisser, T. W. Lee and M. Otsuka, "Speech Enhancement in a Noisy Car Environment," Proc. 3rd International Conference on Independent Component Analysis and Source Separation. pp. 272-277, 2001.

[9] J. F. Cardoso, "Learning in manifolds: the case of source separation," Proc. IEEE SSAP '98, Portland, Oregon.

[10] T. M. Cover, and J. A. Thomas, *Elements of information theory*, New York: Wiley.

[11] A. Hyvarinen, and E. Oja, "Independent component analysis: algorithms and applications," Neural Networks, vol. 13, no. 4/5, pp. 411-430, 2000.

저 자 소 개

김 선 일(정회원)

1983: B.S. degree in electronics engineering, Ajou University, Korea
1985: M.S. degree in electronics engineering, Ajou University, Korea
1996: Ph.D degree in electronics engineering, Ajou University, Korea
1985~1990: Senior research engineer, Korea Institute of Machinery & Metals
1990~Present: Professor, Department of Information Technology for Shipbuilding
1997: Visiting Professor, CAIP Center, Rutgers University, N.J., USA
2007: Visiting Professor, Department of ECE, Georgia Institute of Technology, G.A., USA
<Interested Area : Speech Recognition, Signal Processing>